Lecture 34

In this lecture we continue our study of statistical distributions, looking at expectation and variance.

Averages

The term "average" has various meanings, but most commonly it is taken to be the <u>mean value</u> of a "set" of numbers:

$$\text{mean} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Sometimes, though, another kind of average is meant. This is the <u>median</u> value. Loosely speaking, half of the values are below the median and the other half are above.

More formally, we can list the numbers in increasing order and locate the middle element in the list (if n is odd) or the two middle numbers (otherwise).

In the latter case we take the median to be the mean value of the two middle numbers.

E.g. $X = \{2, 3, 4, 4, 5, 7, 8\}$

$$\text{median} = 4$$

$$\text{mean} = \frac{2+3+4+4+5+7+8}{7}$$

$$= \frac{33}{7} \approx 4.71$$

E.g. $X = \{1, 2, 2, 3, 4, 5, 7, 8, 10, 10\}$

$$\text{median} = \frac{4+5}{2} = 4.5$$

$$\text{mean} = \frac{52}{10} = 5.2$$

Another measure sometimes known as an "average" is the mode, which is the most common value.

In our first example above, the mode was 4 (which happened to coincide with the median). But our second example was "bimodal" in the sense that two values 2 and 10 occurred "equally most often".

In general the mode is less useful than the other two measures.

A single number that in some way is used to characterise a set of data is called a statistic. The mean is the best known and most important statistic.

The mean is known as a measure of central tendency, as is the median. But there's another question we often ask about data.

The question is this. What can be said about the spread of the data? For example, are the numbers all closely clustered about the mean value, or do they spread a long way from the mean?

For example,

$$\{9, 9, 9, 10, 11, 11, 11\}$$

and

$$\{1, 3, 6, 10, 14, 17, 19\}$$

both have mean 10, but the two sets are very different in the way the data is spread about the mean.

Before describing ways to characterise the spread of data, we need to put the discussion in a more general context.

Expectation

Let X be a discrete random variable.

Then the *expected value* or *expectation* of X is

$$E(X) = \sum_x x \Pr(X = x).$$

When X takes only finitely many values $x_1, \ldots, x_n$ then the formula becomes

$$E(X) = \sum_{i=1}^{n} x_i \Pr(X = x_i).$$

The expectation is also called the *mean* (or average) of X, and is also denoted by $\overline{X}$ (or, sometimes, $\overline{x}$).

E.g. (Casting a die)

X	$\Pr(X = x)$	$x \Pr(X = x)$
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6

Then
$$E(X) = \sum_{x=1}^{6} x \Pr(X = x)$$
$$= 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6$$
$$= \frac{1+2+3+4+5+6}{6}$$
$$= \frac{21}{6} = 3.5.$$

E.g. Let's revisit an earlier example, where we had a list

2, 3, 4, 4, 5, 7, 8

of numbers. This example implicitly defines a random variable X with values 2, 3, 4, 5, 7 and 8 where 4 is twice as likely to occur as the other numbers individually are. Then the expectation can be calculated as follows.

X	$Pr(X=x)$	$x\,Pr(X=x)$
2	1/7	2/7
3	1/7	3/7
4	2/7	8/7
5	1/7	5/7
7	1/7	7/7
8	1/7	8/7

Then

$$E(X) = \sum x\,Pr(X=x)$$

$$= \frac{2+3+8+5+7+8}{7}$$

$$= \frac{33}{7} \approx 4.71$$

as before.

Now we are in a position to discuss the variability of data.

On average, how far away are the values of X from their mean? The answer is $E(X-\bar{X})$, which equals 0. So that's not a useful statistic, unless we make $X-\bar{X}$ positive.

This could be done by taking the absolute value $|X-\bar{X}|$, but it's more computationally useful to take the square $(X-\bar{X})^2$ of $X-\bar{X}$.

The statistic $E[(X-\bar{X})^2]$ is called the <u>variance</u> of X:

$$\begin{aligned}
\mathrm{Var}(X) &= E\left[(X-\bar{X})^2\right] \\
&= \sum (x-\bar{X})^2 \Pr(X=x) \\
&= \sum \left[x^2 - 2x\bar{X} + (\bar{X})^2\right] \Pr(X=x) \\
&= \sum x^2 \Pr(X=x) - 2\bar{X} \sum x \Pr(X=x) \\
&\qquad + (\bar{X})^2 \sum \Pr(X=x) \\
&= E(X^2) - 2\bar{X}\,\bar{X} + (\bar{X})^2 \cdot 1 \\
&= E(X^2) - (\bar{X})^2 \\
&= E(X^2) - [E(X)]^2
\end{aligned}$$

E.g. We return to the die-casting example, and put two extra columns in the table.

x	x^2	$\Pr(X=x)$	$x\Pr(X=x)$	$x^2\Pr(X=x)$
1	1	1/6	1/6	1/6
2	4	1/6	2/6	4/6
3	9	1/6	3/6	9/6
4	16	1/6	4/6	16/6
5	25	1/6	5/6	25/6
6	36	1/6	6/6	36/6
			21/6	91/6
			↑	↑
			$E(X)$	$E(X^2)$

$$E(X) = \frac{21}{6} = \frac{7}{2} \ (= 3.5)$$

$$\therefore \text{Var}(X) = E(X^2) - [E(X)]^2$$

$$= 91/6 - (7/2)^2$$

$$= 91/6 - 49/4$$

$$= \frac{182}{12} - \frac{147}{12} = \frac{35}{12} \simeq 2.917.$$